VISUALIZATION OF INTERNAL ARTICULATOR DYNAMICS AND ITS INTELLIGIBILITY IN SYNTHETIC AUDIOVISUAL SPEECH

Katja Grauwinkel, Britta Dewitt, Sascha Fagel

Institute for Speech and Communication, Berlin University of Technology, Germany
Katja.Grauwinkel@tu-berlin.de, brittadewitt@freenet.de, Sascha.Fagel@tu-berlin.de

ABSTRACT

This paper presents the result of a study investigating the influence of visualization of internal articulator movements on the intelligibility of synthesized audiovisual speech. A talking head was supplemented by internal passive and active articulators. A comparative perception test before and after two different training lessons was carried out, where one type of display included all internal articulator movements and the other displayed dynamics without tongue dorsum height, velum opening/closing and tongue forward/backward movements. Results show that recognition scores were significantly higher in audiovisual compared to auditory-alone presentation with non-significantly different recognition scores for both kinds of display. However, only when presenting all additional motion information, the training lesson was able to significantly increase the visual-alone and audiovisual speech intelligibility.

Keywords: talking head, speech visualization, internal articulators, speech intelligibility

1. INTRODUCTION

Many studies have documented that pertinent visual information enhances speech intelligibility when added to audible speech (e.g. [7], [8]). Previous work of the authors could show that this can also be achieved by synthetic audiovisual speech [4]. Computer-based audiovisual speech synthesizers and speech visualization systems provide an interesting tool for investigating bimodal sensory integration, because each parameter of each source of information can easily be manipulated and is therefore experimentally controllable. Furthermore a talking head that might be used as speech trainer (e.g. in foreign language acquisition or speech therapy) offers the possibility to show the internal articulators to explain the production of different speech sounds. Especially the dynamic information of the articulatory displacements of internal articulators are of importance, so that coarticulatory effects can be explained (instead of static states). Therefore a talking head was supplemented by the internal passive and active articulators: alveolar ridge, palatum, velum and pharynx wall. To what extent the dynamic information of these internal articulators is capable to enhance speech intelligibility was evaluated in a second step. If the given visual information of internal articulators dynamics enhances speech intelligibility (either before or after a learning lesson), then the talking head is assumed to be applicable as a tool for speech training and speech therapy.

2. SYSTEM DESCRIPTION

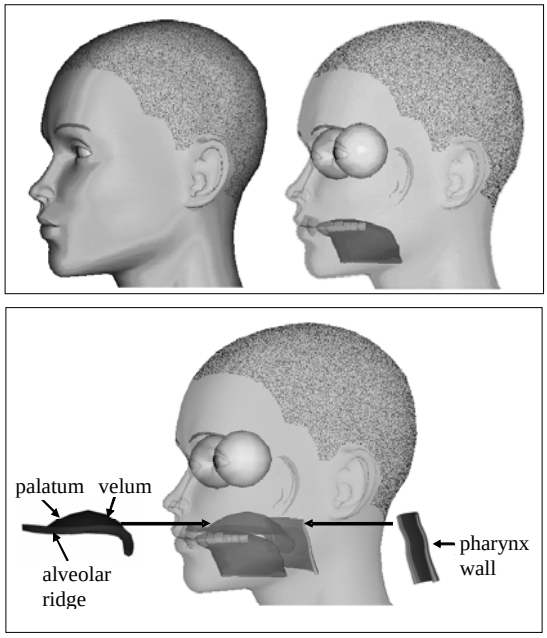
The talking head that was used in the study is a web based modular audiovisual text-to-speech synthesizer with a 3-dimensional animated head [3]. The embedded modules are a phonetic articulation module, an audio synthesis module, a visual articulation module, and a face module. The phonetic articulation module creates phonetic and prosodic information, which consist of an appropriate phone chain, phone and pause durations and a fundamental frequency course. From this data, the audio synthesis module generates the audio signal and the visual articulation module generates motion information. The audio signal and the motion information are merged by the face module to create the complete animation. The MBROLA speech synthesizer [2] is embedded in the audio synthesis module to generate audible speech. The visual articulation module generates a series of parameter settings for the (virtual) articulators. The parameters which were already embedded were:

- tongue tip height,
- tongue dorsum height,
- lip width,
- lip height,
- lower lip retraction and
- lower jaw height.

For each phone and articulator the articulation module provides a coarticulated target position which is held for a fixed fraction of the phone duration. The face module animates the 3D head by linearly displacing its vertices. For each articulator a set of displacement vectors for the head's vertices is defined. Each vertex displacement is a linear combination of the articulators' displacement vectors. The VRML file format (Virtual Reality Modeling Language) is used as output. The facial skin can be displayed opaquely or transparently in order to see the tongue movements.

In this study the talking head's articulators were supplemented by the alveolar ridge, hard and soft palate, uvula, and pharynx wall. For recreating the 3-dimensional geometric models of the internal articulators, the results of medical imaging techniques (in particular mid-sagittal MRI slices) were used as templates. The system configuration before (opaque and transparent) and after changes (transparent) can be seen in figure 1.

Figure 1: Talking head in non-transparent and transparent view before (top) and transparent view after (bottom) supplement of internal articulators.



Additionally to the six existing ones a new articulatory parameter was assigned to the tongue, i.e. backward/forward movement, and another one was defined for the opening/closing of the velum. For these new motion parameters new values for control parameters had to be defined. Because of the refinement of tongue movement new values for tongue dorsum height had to be assigned, too. Consequently the number of control parameters within the visual articulation module increased. The parameter values for the visual articulation module were derived from previous measurement data of electromagnetic articulography [3]. Thus it is based on human articulation movements. The electromagnetic articulography allows to survey articulatory movements at discrete flesh points of the articulators very precisely in space and time, even if they take place inside the mouth. The articulation model implements the dominance principle as suggested by Löfqvist [6] in order to deal with coarticulation.

3. EXPERIMENTAL SETUP

3.1. Items and conditions

The corpus contained ten German consonants with different places of articulation all voiced except the alveolar voiceless fricative [S]; as for its voiced counterpart [Z] no orthographic symbol exists in German and hence could not have been presented to the subjects. The consonants [b,d,g,z,S,v,l,m,n,N] were chosen and combined with the vowels [a,i,u] in a vowel-consonant-vowel (VCV) structure. These vowels were chosen because they define the vertices which span the articulatory/acoustic vowel space. The items were synthesized for the conditions: audio (*A*), visual (*V*), and audiovisual (*AV*). The German female MBROLA voice de7 was selected for creating test stimuli. The audio signal both in *A* and *AV* conditions was embedded in white noise with a signal-to-noise-ratio of 0 dB. On the one hand the visual stimuli in both *V* and *AV* conditions were synthesized with movements of all internal articulators. On the other hand they were synthesized without movements of the internal articulators: backward/forward movement of the tongue, and motion of tongue dorsum and velum were removed. Without these movements only the articulatory movements remain, which can also be observed in a face-to-face conversation.

3.2. Method

20 subjects (from 23 to 59 years old) with normal hearing and normal or corrected-to-normal vision participated in the experiment voluntarily. The subjects had no explicit phonetic knowledge. Stimuli were presented by use of three different quasi-random orders. All the stimuli were presented to the subjects three times before they had to give the answer. The test was designed as a

forced choice test. The recognition of the vowels of the VCV-stimuli were not tested. All subjects were divided into two groups of 10 subjects (group A and B). The stimuli with movements of the internal articulators were presented to group A, the stimuli without these movements were presented to group B. After this pre-test these groups were again divided in two subgroups (group A1, A2, B1 and B2). Groups A1 and B1 received a training lesson of about 30 minutes length with the same motion information as in pre-test, respectively. Afterwards they had a break of 20 minutes. Groups A2 and B2 did not perform the training lesson, they only had a break of the same length. Afterwards a post-test was performed. This test was the same as the one performed as pre-test.

3.2.1. Training Lesson

The training lesson was a video clip in which the articulatory movements for each consonant in each vowel context were explained, while the internal articulators were shown. In order to explain differences in place and manner of articulation pairs of consonants were set in contrast to one another. Articulatory movements of all displayed internal articulators were explained to group A1, whereas the motion information of velum, tongue dorsum and forward/backward displacements of the tongue were not shown to group B1; only lip, jaw and tongue tip movements were explained to them. The articulation process of each consonant-vowel combination was shown in reduced speed while giving explanations. Then the stimuli were shown like those in the test situation but without noise. The training lesson was not interactive, the subjects were told to listen and watch carefully.

4. RESULTS

Analyses of the recognition scores in the pre-test revealed no significant differences in either condition between group A (*A* 40%, *V* 23.7%, *AV* 56%) and group B (*A* 40,7%, *V* 28.4%, *AV* 59%). For statistical significance McNemar's chi-square test was performed. The overall recognition scores in both groups are significantly higher in the *AV* condition compared to *A* ($p<.001$), and *A* scores are significantly higher compared to *V* ($p<.01$).

Closer analyses of the recognition scores with respect to the manner of articulation revealed that even though group A was provided with more motion information compared to group B, there are no significant differences between both groups in the pre-test. Even the nasals, for which the manner of articulation was visualized by lowering the velum for group A, were identified only marginally better by group A in the *V* and *AV* conditions.

Subjects of group A identified stimuli neither significantly better nor worse than group B. Hence subjects were not able to use this additional information without further explanation, but at least they also did not get confused by it. Groups A2 and B2 who did not receive a learning lesson serve as control groups in order to determine if the pre-test must be considered as training.

Table 1 shows the overall recognition scores for each subgroup in each condition. As can be seen, pre-test had no training effect, as recognition scores of A2 and B2 were not significantly different between pre- and post-tests. The increased recognition scores of A1 from pre- to post-test in *V* and *AV* conditions must therefore be due to the learning lesson. Consonants which could not be differentiated visually or audiovisually in pre-test could be identified after the learning lesson.

Table 1: Recognition scores in % for all subgroups in pre- and post-test for each condition.

sub-groups	tests	condition		
		A	***V***	***AV***
A1	pre	40.7	24.7	52.7
	post	44.0	54.7	68.7
A2	pre	40.7	22.7	59.3
	post	46.0	30.7	52.0
B1	pre	44.0	30.7	64.0
	post	41.3	32.7	66.7
B2	pre	36.0	26.0	54.0
	post	38.0	32.0	57.3

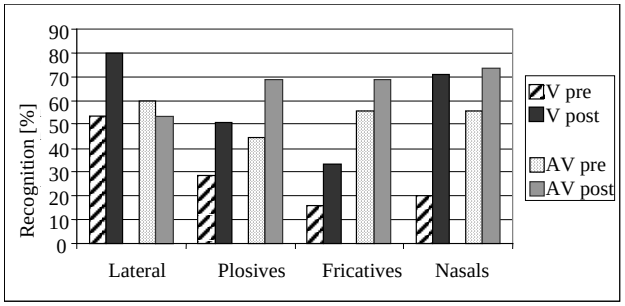
Table 2 displays the recognition scores for groups A1 and A2 in pre- and post-test for conditions *AV* and *V*. Non-significantly different scores are displayed in same columns. As can be seen in the table the learning lesson of group A1 had significant ($p<.001$) influence on *AV* and *V* recognition scores.

Table 2: Recognition scores in % for groups A1 and A2 in pre- and post-test for condition *AV* and *V*. Non-significantly different scores in same columns.

	AV		***V***	
A1 pre	52.7		24.7	
A2 pre	59.3		22.7	
A2 post	52.0		30.7	
A1 post		68.7		54.7

The increase of recognition scores for group A1 between pre- and post-test was higher for *V* (30%) than for *AV* (16%). As can be seen in figure 2, in *V* condition the improvement is mostly due to a better recognition of nasals, in *AV* condition it is mostly due to a better recognition of plosives, furthermore nasals and fricatives.

Figure 2: Recognition scores in % of group A1 in *V* and *AV* conditions in pre- and post-test, respectively, for manner of articulation.



As group B1 only had limited motion information (movements of internal articulators were not shown), the visualizations of the consonants [b,m], [d,n,l], [g,N], [z,S] and [v], respectively, did not differ from one another, which corresponds to the visual distinguishability of consonants from the outside view. Hence, the recognition scores were also analyzed in terms of viseme classes. Nevertheless, neither the recognition scores for consonants nor the recognition scores for viseme classes changed significantly from pre- to post-test in group B1. In contrast to group A1, group B1 was not able to benefit from a training lesson. Without the dynamic information of tongue dorsum height, velum opening/closing and tongue forward/backward movements the recognition scores could not be enhanced significantly.

5. CONCLUSIONS AND FUTURE WORK

On the one hand it could be shown that additional information which is not transmitted in a natural face-to-face communication does not increase speech intelligibility without further explanation. Before a learning lesson subjects were not able to interpret this additional information and hence they were not able to better identify the speech stimuli. However, through a (short) learning lesson a significant enhancement of recognition scores in visual and audiovisual conditions can be achieved.

On the other hand it was shown that a limited amount of visual information of internal articulator movements does not result in such an enhancement. It is assumed that the perception and interpretation of articulatory movements, which can partly be observed in a natural face-to-face communication, cannot be enhanced by a training lesson of 30 minutes.

An important finding is that the given visual information of internal articulator dynamics could be used in order to interpret articulatory speech production. Visual feedback which normally is not available can be learned from subjects with normal hearing and speaking abilities. This study can be regarded as a precursor to an investigation addressing the utility of audiovisual information for speech training. Future work will be dedicated to the question whether people with speech or hearing impairment are also able to interpret and learn from the articulatory dynamics of a talking head. As could be shown in a study by Albert [1] children from 4 to 8 years of age are able to recognize articulatory visual patterns of speech sounds from the visual model of articulation called SpeechTrainer [5]. After a short introduction into the system, which displayed two-dimensional animations of sketches of mid-sagittal MRI slices, the children were able to interpret the information in terms of their own sound production. The authors of the present study currently investigate the applicability of the talking head with its three-dimensional animation of internal articulator dynamics as a method of speech visualization for the use in speech therapy for children with sigmatismus interdentalis.

6. REFERENCES

[1] Albert, S. (2005): Einsatz eines visuellen Artikulationsmodells in der Artikulationstherapie bei Kindern. Diploma thesis at the RWTH Aachen.

[2] Dutoit, T., Bataille, F., Pagel, V., Pierret, N. & van der Vreken, O. (1996): The MBROLA Project: Towards a Set of High-Quality Speech Synthesizers Free of Use for Non-Commercial Purposes. *Proc. ICSLP*, 1393-1396.

[3] Fagel, S., Clemens, C. (2004): An Articulation Model for Audiovisual Speech Synthesis - Determination, Adjustment, Evaluation. *Speech Communication* 44: 141-154.

[4] Grauwinkel, K. & Fagel, S. (2006): Crossmodal Integration and McGurk-Effect in Synthetic Audiovisual Speech. *Proc. SPECOM*, St. Petersburg, Russia.

[5] Kröger, B.J. (1998): *Ein phonetisches Modell der Sprachproduktion*. Niemeyer Verlag, Tübingen.

[6] Löfqvist, A. (1990): Speech as Audible Gestures. In W. J. Hardcastle, A. Marchal (eds.), *Speech Production and Speech Modeling*, Dodrecht: Kluwer.

[7] Sumby, W.H., Pollack, I. (1954): Visual Contribution to Speech Intelligibility in Noise. *JASA* 26, 212-215.

[8] Summerfield, Q. (1979): Use of Visual Information for Phonetic Perception. *Phonetica* 36: 314-331.